7.2 DEVELOPMENT OF A 103-YEAR HIGH-RESOLUTION CLIMATE DATA SET FOR THE CONTERMINOUS UNITED STATES

Wayne Gibson[1*], Christopher Daly[1], Tim Kittel[2], Doug Nychka[2], Craig Johns[2], Nan Rosenbloom[2], Alan McNab[3], and George Taylor[1]

[1] Oregon State University, Corvallis, OR 97331, USA
[2] National Center for Atmospheric Research, Boulder, CO 80307, USA
[3] National Climatic Data Center, Asheville, NC 28801, USA

1. INTRODUCTION

Currently, the only high-quality, high-resolution temperature and precipitation data sets for the continental United States suitable for use on climatological time scales are for mean values. None yet exist that represent sequential monthly values over an extended historical period. Such data sets would enable, for example: transient ecological, hydrological, and natural resource modeling for use in global change assessment; analysis of local and regional climate variations; analysis of frequency, duration, and spatial patterns of extreme climatological events; and investigation of relationships between climatological variability and large-scale forcing mechanisms (e.g., ENSO or QBO).

A small number of gridded sequential data sets are under development, but they are at a coarse resolution or have a limited spatial extent. The only long-term sequential data sets currently available for the U.S. are for individual sites or are based on simple averages of any available stations within major climatic divisions. The site data are sparse and located primarily in valley bottoms, and the climatic division data suffer from inherent biases in sensor placement and availability over time.

The current project, under NOAA/NASA funding, involves development of 103 years (1895-1997) of gridded monthly precipitation and maximum and minimum temperature at a 4-km resolution for the contiguous United States. PRISM, a knowledge-based interpolation system was used to produce the gridded climate datsets (Daly et al. 1994, 1997, 2002). The project involves development of serially complete monthly data for approximately 8,000 stations in the U.S. An important part of this work has been to develop and apply a semi-analytical, spatially-intelligent quality control system, based on PRISM, for monthly station observations.

* *Corresponding author address:* Wayne Gibson, Oregon Climate Service, Oregon State University, 316 Strand Agricultural Hall, Corvallis, OR 97331; e-mail: gibson@oce.orst.edu

2. PROJECT OVERVIEW

The main objective of this project is to produce a high-quality, topographically-sensitive, 103-year data set of monthly temperature and precipitation on a 2.5-min (4-km) grid over the conterminous United States (Daly et al 1999).

This has been a four-year project. In Year 1, preliminary precipitation grids for 1948-93 were generated. Based on these results, the schedule of tasks in Year 2 was organized to reflect the emergence of scientific issues surrounding data set quality and usefulness. As we were taking a close look at the station data and refining our techniques for infilling missing months, we began to ask serious questions about the quality of the data. In the end, we found it most efficient to address issues of data quality up-front, rather than waiting until the final gridded products were completed. As a result, a significant amount of work went into the development and application of a sophisticated, spatial QC system for station data.

The QC system involved the following aspects:

Task 1: Extend the time period of the data set to 103 years by adding data through 1997, and add station data for several hundred pre-1948 stations recently digitized by the Midwestern Climate Center.

Task 2: Develop and apply a semi-automated, spatially intelligent quality control system for monthly precipitation observations.

2.1 *Task 1 - Expand Collection Of Station Data*

In order to make the monthly data sets as useful and timely as possible, coop station data for 1997 were added to the data base. This brings the total length of the data set to 103 years: 1895-1997. The additional data were gathered at NCDC, transferred to NCAR for reformatting, then passed to OSU for quality checking.

It is well-known that station data are very sparse in the early part of this century, especially before 1948, when observations are not routinely available in digital form. Sparse data will limit the usefulness of spatial data sets for natural resource modeling and climate trend and variability analyses, because the

inter-annual variability in data-sparse areas is highly smoothed spatially. The Midwestern Climate Center recently completed the digitizing of pre-1948 daily temperature and precipitation data for several hundred stations in the midwestern region (plus New Mexico). Although there was a significant investment of time associated with our assimilation of the daily data and the subsequent production of monthly data, we felt that including the MCC stations increased the usefulness of the data set significantly.

Our station data set now consists of the following classes of data:

COOP: NWS cooperative stations with sequence numbers that represent station location and/or instrumentation changes

SNOTEL: NRCS Snow Telemetry data. Begins in 1978.

AG: Agricultural climate data for southeastern Washington. From NRCS.

MCC: Midwest Climate Data Center data. Data recently digitized. Mainly for pre-1948 period.

HCN: Historical Climate Network data (supercedes COOP data for the same ID code).

MISC: Miscellaneous data, including storage gauges, snow courses, manually estimated points.

2.2 *Task 2 - Develop A PRISM-based QC System For Monthly Station Data*

Spatial modeling and mapping of climatic observations is a relatively new science, and is placing higher demands on data quality than traditional, point-based analyses. Previous experience with spatial modeling projects has demonstrated the problems experienced with insufficiently quality-controlled (QC) data sets. A single outlier has the capability of causing spatial abnormalities, sometimes quite large and obvious. The outlier may not appear unusual in a time series plot for a single station, but spatial analyses can often show the inconsistency of the outlier in relation to values from nearby stations.

While a spatial analysis can be an effective way of identifying outliers, the detection and correction of such errors after the spatial data set has been produced can be time-consuming and inconsistent. Each spatial grid must be visually examined for possible outliers, and special calculations must be made to help show outliers where they may be otherwise hidden, such as in areas of complex terrain. In this project, there are 1236 monthly grids, each covering the entire lower 48 states. Examination of such a large number of grids requires many man-weeks of time, and is fraught with inconsistencies inherent in human visual inspection. When possible errors are found, they must be individually verified, and the erroneous grid reproduced without the offending data point.

Our premise is that it is better to use the power of the spatial analysis system (i.e., PRISM) *a priori* to detect errors in the station data in as an objective and automated fashion as possible. We now have produced a semi-automated system that is extremely effective at identifying bad monthly data, while leaving good data untouched. Our definition of "bad data" is a value that has experienced an error in transcription somewhere between the observation and the digital value available to us. Here, we are not attempting to identify errors due to gauge undercatch, or subtle inhomogeneities cause by changes in observation method, instrumentation changes, etc.

The QC system incorporates two main types of checks:

(1) Metadata errors - errors in the location or elevation of a station. Such errors are insidious, in that a station mislocation affects the entire period covered by the erroneous metadata, not just one month.

(2) Monthly data errors - errors in the actual monthly data values. This task was conducted using "ASSAY_QC"

ASSAY_QC is a version of PRISM, the knowledge-based interpolation system used to produce the gridded climate datsets (Daly et al. 1994, 1997, 2002). ASSAY_QC is a modified version of ASSAY, which is essentially a full version of PRISM, except that it makes predictions for individual points in space, rather than producing gridded output. ASSAY is commonly used to produce jackknife cross-validation errors for various model parameterizations. For use in QC, ASSAY_QC made a prediction for each station for each month, in the absence of that station's data value. Large differences between the predicted and observed values indicated that there was a discrepancy between the station observation and those from surrounding stations. To determine what constitutes a "large" discrepancy or a consistent basis, we derived a strong empirical relationship between precipitation and variance from climatology; the higher the precipitation, the larger the expected variance. This relationship was used to normalize differences between predicted and observed precipitation.

Data outliers were identified based on the following set of criteria:

1. Distance Confidence Criterion. We have less confidence in the ASSAY_QC predictions the farther away the surrounding stations are from the predictor location.
2. Elevation Confidence Criterion. We have less confidence in the ASSAY_QC predictions the larger the elevation differences between the surrounding stations and the predictor location.
3. Precipitation Confidence Criterion. We have low confidence in our ability to detect outliers when there is a small difference between the observation and the prediction, as is often the case when the observed precipitation is low.

4. Prediction vs Observation Criterion. Does a "large" difference exist between what we expect and what was modeled?

Station observations that satisfied all of these criteria were considered candidate outliers. Using additional software, they were then divided into several groups, including definite outliers, those that appeared to be outliers only because the most highly weighted station was also an outlier, cases that were too close to call, and others. Those that were too close to call were examined manually. Those classified as definite outliers were set to missing.

The QC system detected 2,371 monthly data errors out of 6,345,675 station-months, for an error detection rate of 0.0374%. While this may appear to be a very low error rate, it translates into about two errors per monthly grid, which is not insignificant. Fifty-six outliers were too close to call; after manual checks, 40 were found to be erroneous and 16 could not be proven invalid. This reasonable split between good and bad data in a group that was deemed marginal confirmed that the system was operating as an effective substitute for an expert operator.

3. CONCLUSIONS

High-resolution, 103-year, sequential monthly data sets of temperature and precipitation are being constructed for the conterminous United States. These data sets are unprecedented in their combination of high quality, high resolution, and century-long duration. They will be used to support numerous modeling and analysis activities, such as ecological modeling, trends and variability analyses, extreme event analyses, and investigation of relationships between climatological variability and large-scale forcing mechanisms.

Final precipitation and minimum and maximum temperature grids for the full 1895-1997 period are expected to be available in summer 2002.

4. REFERENCES

Daly, C., W. P. Gibson, G.H. Taylor, G. L. Johnson, P. Pasteris. 2002. A knowledge-based approach to the statistical mapping of climate. *Climate Research*, in press.

Daly, C., T.G.F. Kittel, A. McNab, J.A. Royle, W.P. Gibson, T. Parzybok, N. Rosenbloom, G.H. Taylor, and H. Fisher. 1999. Development of a 102-year high-resolution climate data set for the conterminous United States. In: *Proceedings, 10th Symposium on Global Change Studies*, 79th Annual Meeting of the American Meteorological Society, 10-15 January, Dallas, TX, 480-483.

Daly, C., R.P. Neilson, and D.L. Phillips. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology* 33: 140-158.

Daly, C., G.H. Taylor, and W.P. Gibson. 1997. The PRISM approach to mapping precipitation and temperature. In: *Proc., 10th AMS Conf. on Applied Climatology,* Amer. Meteorological Soc., Reno, NV, Oct. 20-23, 10-12.

USDA-NRCS. 1998. *PRISM Climate Mapping Project--Precipitation. Mean monthly and annual precipitation digital files for the continental U.S.* USDA-NRCS National Cartography and Geospatial Center, Ft. Worth TX. December, CD-ROM.